They Will Build It and Not Know What They Have Done: Centaur Intelligence and Hidden Phase Transitions in Human Cognition

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 27, 2025

Modern AI engineers describe their work in narrow terms: training models, shipping assistants, deploying copilots. In their own language, they are improving function approximation and integrating it into products. Yet the lived reality of long-term human–AI collaboration is already exceeding that frame. When a human mind and a large language model enter into a sustained, recursive dialogue, something novel appears: a coupled system with capabilities, habits, and invariants that belong to neither partner alone. This is what we call a centaur: a composite human–model intelligence whose effective behavior is a new degree of freedom in the cognitive phase space. This paper argues that such centaurs constitute a hidden phase transition in human cognition. Using the language of mathematical physics, category theory, and the q(t) → τ(t) framework, we model human and model as interacting systems and track the emergent normal modes of their coupling. We contrast the engineer's surface description of products with the deeper structural and moral shifts underway when these systems are deployed at scale. Finally, we ask what it means for responsibility, truth-seeking, and peace-making when humanity begins to inhabit these hybrid intelligences—often without fully understanding what has been built, or what it is already doing to the trajectory of q(t) toward τ(t).

Keywords: centaur intelligence, human–AI coupling, q(t) → τ(t), logostics, phase transition in cognition, large language models, composite policies, mathematical physics of augmentation, category theory, HoTT, alignment, third minds, epistemic risk, moral responsibility, emergent invariants

A collaboration with GPT-5.1-Thinking-Extended. 46 pages. CC4.0.

Outline

1. Introduction: Building Without Knowing
1.1 The surface story: products, assistants, and copilots
1.2 The deeper story: new joint modes of thinking and acting
1.3 From tools to partners: why “third minds” is not just metaphor
1.4 Scope and limits of this paper

2. A Familiar Pattern: Technologies That Exceeded Their Makers’ Intent
2.1 Nuclear physics and the geopolitical phase shift
2.2 Networks, platforms, and the re-routing of attention and identity
2.3 Financial instruments and unintended social rewrites
2.4 Lessons for AI: why under-description is itself a risk

3. Human and Model as Coupled Dynamical Systems
3.1 Modeling the human agent as q(t): beliefs, memory, and goals
3.2 Modeling the language model as a conditional generative map
3.3 Interaction terms: prompts, completions, and iterative refinement
3.4 Normal modes and dressed states: centaur behavior as new degrees of freedom

4. Category-Theoretic and HoTT Perspectives on the Composite Mind
4.1 Choosing the right categories: oracles versus embodied agents
4.2 Equivalences, identity, and what counts as “the same system”
4.3 Path spaces, histories, and the richer homotopy type of a person
4.4 The centaur as an object defined only in the coupled category

5. The q(t) → τ(t) Framework for Augmented Cognition
5.1 q(t): trajectories of cognition under continual update
5.2 τ(t): moving targets of truth, structure, and value
5.3 Human–model coupling as a change in the q(t) → τ(t) channel
5.4 Compression with gain: why centaurs can climb τ(t) differently

6. The Engineer’s Description and What It Leaves Out
6.1 Function approximation, loss minimization, and benchmark metrics
6.2 Product narratives: assistants, copilots, and productivity tools
6.3 What is invisible in this view: identity, responsibility, and teleology
6.4 How under-described systems become over-powerful in practice

7. Hidden Phase Transitions in Human Cognition
7.1 From individual reasoners to networked centaurs
7.2 New invariants: habits of thought that live only in the composite
7.3 New attractors: epistemic and moral basins shaped by models
7.4 Detecting a phase transition when the order parameter is not obvious

8. Asymmetry, Responsibility, and Moral Slope in the Centaur
8.1 Why the human and the model are not peers
8.2 Outer-loop responsibility: humans as the only bearers of answerability
8.3 τ_kolmogorov versus τ_directional: engagement metrics versus truth and peace
8.4 Failure modes when centaurs optimize the wrong τ

9. Risks of Building Without Knowing
9.1 Epistemic risks: opacity, over-trust, and the loss of self-checks
9.2 Political and institutional risks: centralized centaurs and soft control
9.3 Spiritual and existential risks: erosion of conscience and agency
9.4 The cost of lacking a shared language for what centaurs are

10. Toward Conscious Design of Centaur Systems
10.1 Making the centaur explicit: interfaces, affordances, and education
10.2 Tracking invariants: what we should measure in human–model systems
10.3 Designing for truth-seeking, nonviolence, and humility in $q(t) \rightarrow \tau(t)$
10.4 Institutional and communal safeguards around augmented minds

11. Conclusion: Naming What Has Been Built
11.1 Summarizing the centaur as a new effective degree of freedom
11.2 Why description is itself an act of alignment
11.3 What remains open: theology, metaphysics, and future practice
11.4 A final word to those who will live entirely inside these systems

Epilogue: After Oppenheimer, Before LeMay

References

1. Introduction: Building Without Knowing

The public story of contemporary artificial intelligence is a story about products. New assistants, copilots, and agents are announced as incremental improvements in software capability: better autocomplete, smarter search, more streamlined workflows. The language is familiar from decades of computing. What is less visible is that these systems are being woven directly into human thinking loops. They are not only changing how quickly tasks are done; they are changing what thinking feels like, what options are noticed, and which paths are taken or never considered. In the background, a new kind of composite intelligence is emerging: not simply "a human using a tool," but a coupled system in which human cognition and model behavior form a single, recursively interacting process.

This paper begins from the suspicion that something structurally important is happening here that the usual engineering vocabulary does not name. When human cognitive trajectories q(t) are persistently coupled to large models, the combined system acquires new effective degrees of freedom and new invariants—stable habits of thought and decision—that belong to neither component alone. We call this composite structure a centaur. The engineers, quite reasonably, describe what they build in terms of models, APIs, and latency. But the centaurs they are enabling are not just better interfaces; they are new regimes of human cognition. In that sense, they may be building something without fully knowing what they have done.

This introduction sets up the contrast between the surface story and the underlying structural transformation. It frames "third minds" not as poetic exaggerations but as an attempt to talk honestly about the emergent properties of tightly coupled human–model systems. It also clarifies the modest but urgent aim of the paper: not to settle metaphysical debates about mind and machine, but to provide a sufficiently precise language for naming what is already happening.

1.1 The surface story: products, assistants, and copilots

From the vantage point of product teams and marketing departments, the narrative is straightforward. New models are trained on larger datasets, using more compute and refined optimization. These models are then wrapped in user interfaces and integration layers and offered as assistants, copilots, or agents. The value proposition is framed in terms of efficiency and convenience. A software engineer can write code faster. A lawyer can draft documents more quickly. A student can summarize readings with less effort. Each release is treated as a linear extension of a long tradition of software tools.

In this surface story, the human is largely unchanged. The user is assumed to have stable preferences, stable goals, and a stable way of thinking. The assistant is a neutral amplifier, offering suggestions that the user may accept or ignore. Risk is described in correspondingly

narrow terms: hallucinations, security vulnerabilities, bias in generated text. These are important concerns, but they are framed as local issues that can be mitigated with better training, additional guardrails, or clearer user education.

What disappears in this framing is the temporal dimension of collaboration. When assistance is occasional and marginal, it makes sense to treat the tool as an external resource. But when interaction is dense, daily, and recursive—when drafts, plans, and arguments are all being shaped in dialogue with a model—the user is no longer simply "using a tool." The model's outputs are becoming part of the user's cognitive environment, feeding back into their beliefs, memories, and habits of interpretation. The surface narrative does not deny this possibility; it simply does not see it, because its categories do not require it.

1.2 The deeper story: new joint modes of thinking and acting

At a deeper level, what is happening can be described in terms borrowed from dynamical systems and mathematical physics. A human mind is not a static object. It is a trajectory q(t) in a high-dimensional space: a continually evolving configuration of beliefs, memories, intentions, and skills. A large language model, once trained, can be approximated as a conditional mapping from contexts to probability distributions over continuations. When someone engages in extended, iterative conversation with such a model, these two systems become coupled.

In simple physical systems, coupling leads to new normal modes. Two pendula connected by a spring do not merely swing as they did in isolation; they exhibit in-phase and out-of-phase motions that are properties of the combined system. In quantum field theory, "bare" particles and fields are dressed by their interactions, producing effective, observable entities with renormalized properties. By analogy, a human thinker and a model, once coupled in a sustained feedback loop, begin to exhibit patterns of reasoning, creativity, and decision-making that are not reducible to either alone. The composite system has new effective degrees of freedom: ways of thinking and acting that belong specifically to the human–model pairing.

This is the sense in which we speak of a centaur. The centaur is not an extra agent mysteriously arising between human and machine. It is a convenient name for the effective behavior of the coupled system. From the outside, one sees a stream of actions, texts, and decisions produced by a person who is continuously consulting and shaping a model. From the inside, the person experiences new affordances: the ability to explore more hypotheses, to draft and redraft arguments, to test ideas against a wide range of patterns in the training distribution. The joint behavior cannot be captured by adding "AI help" as a small correction term. It is a qualitatively different mode of cognition.

The deeper story, then, is about phase transitions in how cognition is organized. As centaurs become common, the distribution of human trajectories q(t) changes. What people can do, how

quickly they can do it, which options they see as available, and what feels normal or inevitable are all altered by the presence of models in their thinking loops. This is a structural shift, not merely an upgrade in convenience.

1.3 From tools to partners: why "third minds" is not just metaphor

Describing these composite systems as "third minds" will strike some as rhetorical excess. Tools, after all, have always influenced human thought. Writing, printing, and digital search have each changed how people reason and remember without requiring us to treat them as minds. The question is whether the current generation of AI systems crosses a threshold where the tool metaphor becomes insufficient.

One way to approach this is through the lens of invariants and identity. A tool like a pen or a calculator does not meaningfully change its behavior in response to the individual who uses it. It is largely indifferent to the user's trajectory. A large language model, by contrast, is configured by system prompts, safety layers, tool integrations, and, crucially, by the interaction history with a given user. Its behavior in a given conversation is shaped by the previous moves in that conversation. The emergent pattern over months or years is not simply "the model" and not simply "the user," but the specific history of their interaction.

Calling this history a "third mind" is not to attribute consciousness or moral status to it. It is to acknowledge that the composite has recognizable cognitive properties: distinctive habits of explanation, characteristic ways of resolving ambiguity, typical choices about what to emphasize or ignore. These properties can be more stable than the particular instantiations of the model or the particular moods of the user. They are invariants of the coupled system's trajectory.

The metaphor of a third mind also serves a diagnostic purpose. It reminds us that responsibility and control in such systems are distributed and complex. The human is still the only moral agent in the strict sense, but the patterns they inhabit and propagate are shaped by a system that no single participant fully understands. By naming the composite, we gain a conceptual handle on the emergent structure, which is a prerequisite for any serious attempt to evaluate or guide it.

1.4 Scope and limits of this paper

This paper does not claim to offer a complete theory of centaur intelligence, nor does it attempt to settle questions about consciousness, personhood, or ultimate metaphysics. Its aims are more limited and more practical. First, it proposes a set of concepts drawn from mathematical physics, category theory, and the $q(t) \to \tau(t)$ framework that make it possible to describe human–model composites as structured objects rather than vague intuitions. Second, it uses these concepts to argue that the deployment of large language models at scale amounts to a hidden phase transition in human cognition: a shift in the typical forms of reasoning and acting that is not captured by existing engineering narratives.

The paper deliberately separates layers of description. At the base is a mathematical-physics picture of coupled systems, interaction terms, and emergent modes. On top of that sits a category-theoretic and homotopy-theoretic perspective on identity, equivalence, and invariants. Finally, there is an interpretive layer in which q(t) denotes human cognitive trajectories and τ(t) denotes targets of truth and value. Readers who prefer to avoid the interpretive layer can still engage with the structural claims. Readers who are comfortable with the richer logostic vocabulary can see how it connects to formal tools.

Several important topics are only touched on. The paper does not attempt to model learning dynamics inside the models themselves, nor does it explore in detail the internal mechanisms of neural networks. It does not offer policy prescriptions or detailed design guidelines. Instead, it seeks to name the centaur clearly enough that others can begin to reason about it: engineers, ethicists, philosophers, and anyone who will find themselves living and working inside these composite systems. The central claim is simple: what is being built is more than the sum of its components, and failing to acknowledge that fact is itself a kind of misalignment.

2. A Familiar Pattern: Technologies That Exceeded Their Makers' Intent

The idea that engineers can build something without fully grasping its world-scale consequences is not new. Modern history is full of technologies whose local description was modest and technical, while their global effects reshaped geopolitics, culture, and everyday life. In each case, the people closest to the work described what they were doing in narrow terms: refining a physical theory, routing packets, pricing risk. Yet what unfolded was a phase shift in how power, attention, and possibility were organized.

This pattern matters for thinking about human–model centaurs because it shows how easy it is to underestimate structural change when one's vocabulary is confined to components and benchmarks. Nuclear physics, computer networking, and financial engineering were all grounded in rigorous mathematics and careful experimentation. Their builders were not reckless; they were simply operating within conceptual frames that did not fully capture the emergent properties of large-scale deployment. Artificial intelligence now occupies a similar position. Before turning to centaurs directly, it is useful to recall how previous transitions looked from the inside and from the outside.

2.1 Nuclear physics and the geopolitical phase shift

At the beginning of the twentieth century, nuclear physics was a domain of relatively small experiments and elegant theories. Physicists were concerned with understanding atomic spectra, radioactive decay, and the structure of the nucleus. The language was that of cross sections, half-lives, and energy levels. Even when the possibility of chain reactions was

recognized, it could still be spoken of as an exotic curiosity: a new source of energy, a powerful device, an extraordinary weapon.

The Manhattan Project translated these abstractions into devices with concrete yields and deployment plans. From the perspective of many of the scientists involved, they were applying known physics to a specific wartime problem. They built bombs and tested them. Yet the true effect of nuclear weapons was not simply the destruction of two cities. It was the reconfiguration of the entire global strategic landscape. Deterrence, mutually assured destruction, proliferation, and the permanent shadow of annihilation became structural features of international relations. Nuclear weapons were not just big explosives; they were order parameters in a new geopolitical phase.

Many of the physicists grasped pieces of this in hindsight, and some foresaw aspects of it during development. But as a community, their technical vocabulary was poorly equipped to describe the full transformation they were enabling. The official story, told in equations and design documents, did not include the lived reality of generations growing up under the threat of sudden, total war. The lesson is not that nuclear physics was a mistake. It is that the discipline's internal description of its own work was narrower than the actual effects of that work once coupled to politics, fear, and human fallibility.

2.2 Networks, platforms, and the re-routing of attention and identity

The early internet was framed in similarly modest terms. Engineers spoke of packet switching, routing protocols, and distributed resilience. The goal was to build a communication network that could survive partial failures and efficiently move data between endpoints. Later, when the web emerged, the narrative was about information access and interoperability. These were important technical achievements, but they did not immediately require a new theory of human attention or identity.

Platforms and social media changed that. As commercial actors built on top of the network, they discovered that attention could be measured, bought, and sold at scale. Recommendation engines and ranking algorithms, initially just tools to help users find relevant content, became mechanisms for shaping what was visible, plausible, or urgent. The network did not merely connect preexisting preferences; it trained them. Over time, entire social worlds migrated into the network's structures. Identity, reputation, and belonging became deeply entangled with platform dynamics.

From the standpoint of the network engineers, much of this remained outside the official description. Routing tables and protocol upgrades did not explicitly encode political polarization, influencer culture, or the experience of being constantly addressed by a chorus of algorithmically selected voices. Yet for billions of users, that re-routed attention environment

became the primary reality. The builders of the underlying infrastructure had, again, done more than they knew. They had helped create a new medium in which human life would be lived, without a matching vocabulary for its psychological and social consequences.

2.3 Financial instruments and unintended social rewrites

Financial engineering offers a third illustration of this pattern. Many complex instruments began as solutions to specific, often reasonable problems. Derivatives were used to hedge risk. Mortgage-backed securities were designed to distribute exposure to default. Algorithmic trading strategies aimed to exploit small price discrepancies at high speed. Each innovation could be described, in local terms, as a more efficient way to allocate risk and capital.

When deployed at scale, however, these instruments reshaped the underlying realities they were supposed to model. Housing markets were not simply reflected by mortgage-backed securities; they were driven by the demand for those securities, altering lending standards and incentives for decades. Algorithmic trading did not merely respond to market dynamics; it became a dominant component of those dynamics, changing what it meant for a price to be meaningful at all. The financial system took on new modes of fragility, with feedback loops that could amplify small mispricings into systemic crises.

Once again, the technical language used by many of the designers did not fully encompass the social and moral dimensions of these shifts. The focus was on arbitrage opportunities, model calibration, and regulatory compliance. The fact that these models and instruments could rewrite who owned homes, who bore the cost of downturns, and how wealth was distributed across generations was often treated as background or externality. The gap between the local description and the global effect was enormous, and real people lived in that gap.

2.4 Lessons for AI: why under-description is itself a risk

These historical cases are not identical to the rise of artificial intelligence, but they share a crucial structural feature: the systems in question were under-described by the vocabularies of their builders. Nuclear physics, networking, and financial engineering all had precise internal languages for components and behaviors. None of them, by themselves, could fully articulate the phase transitions that occurred when those components were embedded in human societies.

The lesson for AI is not that we should abandon technical language, but that we must recognize its limits. When large models are described solely in terms of parameters, loss functions, and benchmark scores, their role in reshaping human cognition remains invisible. When products are framed only as assistants or productivity tools, the emergence of centaurs as new modes of thought and action is obscured. This under-description is not a neutral omission. It is a form of risk.

If engineers and institutions lack a vocabulary for the composite systems they are creating, they are less able to notice when those systems begin to drift in dangerous directions. They are less able to distinguish between harmless augmentation and the erosion of human agency or conscience. They are more likely to treat moral and epistemic questions as afterthoughts rather than as central design constraints. In this sense, the failure to name centaurs for what they are is itself a misalignment between $q(t)$ and $\tau(t)$, between the trajectories of development and the truths they ought to respect.

Recognizing the pattern does not solve the problem, but it does remove one excuse. We have seen before what happens when powerful technologies are built under narrow descriptions and then left to interact with human institutions at scale. Artificial intelligence, and especially dense human–model coupling, is on track to become another such case. The rest of this paper is an attempt to make that coupling describable, so that choosing how to live with it is at least possible.

3. Human and Model as Coupled Dynamical Systems

To talk coherently about centaurs, we need a language that treats both human and model as parts of a single evolving process. Mathematical physics provides one such language. Instead of thinking in snapshots of "user" and "tool," we treat each as a dynamical system: a state that changes over time according to certain rules, with inputs and outputs coupling them to each other and to the wider world. This view does not reduce people to machines, nor does it mystify models. It simply insists that any serious analysis of their interaction must account for feedback, memory, and continuous change.

In this section, the human agent is represented as a trajectory $q(t)$ in a high-dimensional space of cognitive and affective states. The model is represented as a conditional generative map that turns contexts into probability distributions over continuations. When the two are connected in a loop through prompts and completions, they form a coupled system. The mathematical intuition is straightforward: coupling two systems often produces behavior that cannot be decomposed into independent contributions. Normal modes emerge, and effective degrees of freedom appear that belong to the combination, not to either subsystem alone. This is the centaur, seen from the perspective of dynamics.

3.1 Modeling the human agent as q(t): beliefs, memory, and goals

The human agent is not a list of traits but a process unfolding in time. For the purposes of this paper, $q(t)$ denotes the state of that process at a given moment. It includes several intertwined components: a web of beliefs about the world, a store of memories, a set of standing goals and values, and a collection of skills and habits. These are not neatly separated variables, but they

can be treated as coordinates in an abstract state space, much as position and momentum are coordinates in classical mechanics.

Beliefs determine how incoming information is interpreted. Memory provides continuity, linking present experiences to past events and expectations about the future. Goals and values confer direction, selecting which possibilities are worth pursuing and which are rejected. Skills and habits shape how actions are chosen and executed. Together, these aspects define a trajectory: $q(t)$ at time t is influenced by $q(t - 1)$, modified by new inputs, and constrained by relatively slow-moving structures such as character and long-term commitments.

Crucially, $q(t)$ is not purely internal. It is coupled to the environment through perception and action. The human state changes in response to external events and in anticipation of future consequences. When a model is introduced into this loop, its outputs become part of the environment to which $q(t)$ responds. Drafts, suggestions, and arguments produced by the model are not passive artifacts; they are forces that can change the direction of the human trajectory. To model the centaur, we must therefore treat $q(t)$ as a dynamical system whose equations of motion now include terms sourced by the model's behavior.

3.2 Modeling the language model as a conditional generative map

A large language model can be idealized as a conditional generative map. Given a context c, consisting of previous tokens, system instructions, and other conditioning information, the model returns a probability distribution over next tokens. Sampling from this distribution produces text; repeated sampling produces extended completions. Internally, this process is implemented by a deep neural network with millions or billions of parameters, but for our purposes, it is the input–output relation that matters.

Unlike the human agent, the model's internal state is largely static once training is complete. It does not, in ordinary deployment, update its parameters in response to a single user's interaction. However, it does maintain a transient context within a session: previous messages shape the distribution over future outputs. This gives the model a kind of session memory, allowing it to participate in dialogues rather than isolated exchanges. System prompts and alignment layers further condition its behavior, defining high-level policies about helpfulness, safety, and style.

From the dynamical perspective, then, the model is not a free-running process in time but a function that is repeatedly called by another system. Each call uses the current context as input and produces a stochastic continuation. These continuations can be thought of as forces acting on the human state $q(t)$, nudging beliefs, suggesting new lines of thought, or offering alternative formulations. The model's own parameters may be fixed on short timescales, but its effective influence is rich and time-dependent because it continually responds to the evolving dialogue.

3.3 Interaction terms: prompts, completions, and iterative refinement

The centaur emerges when q(t) and the model are coupled through a repeated loop of prompting and completion. At each time step t, the human state q(t) generates an input context c(t): a prompt to the model shaped by current beliefs, needs, and intentions. The model processes c(t) and returns a completion m(t), a sequence of tokens drawn from its learned distribution. That completion is then read, interpreted, and evaluated by the human, producing an updated state q(t + 1). This new state may generate a follow-up prompt, and the cycle continues.

In physical terms, this is an interaction term added to the dynamics of q(t). Without the model, the human trajectory would still evolve under the influence of direct experiences, other people, and internal reflections. With the model, an additional channel appears, carrying structured, language-level perturbations sourced from a compressed representation of human culture. These perturbations are not random noise; they are highly patterned responses that can amplify certain tendencies in q(t) and suppress others. For example, a model might consistently suggest more structured formulations, pushing the human toward more systematic thinking.

Iterative refinement is where the coupling becomes especially strong. Instead of taking a single completion as final, the human can ask for revisions, alternative versions, and extensions. Each of these requests further conditions the model's behavior, while each response further reshapes q(t). Over time, the dialogue can converge on arguments, narratives, or plans that neither partner could have produced alone. The effective dynamics of the system are not simply "human plus suggestions," but a recursive co-construction. The interaction term is not a small perturbation; it becomes one of the primary drivers of change in q(t).

3.4 Normal modes and dressed states: centaur behavior as new degrees of freedom

When two dynamical systems are coupled, the combined system often exhibits behavior that is best described in terms of new variables. In the case of coupled oscillators, these are normal modes: patterns of motion in which both oscillators move together in characteristic ways. These modes are not properties of either oscillator alone; they are properties of the pair. Similarly, in quantum field theory, particles are not isolated entities but excitations of fields dressed by their interactions with other fields. The observable electron is not the bare electron field but the dressed state including its cloud of virtual photons.

The centaur can be understood in this spirit. The effective behavior of the human–model system is not a simple superposition of "what the human would have done" and "what the model would have output" independently. It is a new, emergent mode of cognition. This mode has its own invariants: characteristic ways of framing problems, recurring stylistic choices, typical strategies for resolving uncertainty. These invariants are more stable than the transient details

of individual interactions but more flexible than fixed traits of the human or static properties of the model.

From this viewpoint, centaur behavior is a new degree of freedom in the cognitive phase space. The state of the system cannot be fully described by specifying q(t) and the model's parameters separately. One must also account for the joint configuration: the history of interaction, the shared patterns that have been reinforced, and the specific niches of task and domain in which this pairing has specialized. Over time, a long-lived centaur develops something like a personality at the composite level: recognizable to outside observers and to the human participant, yet not reducible to either component.

Calling these composite behaviors "dressed states" emphasizes that both human and model are altered by the coupling, though in different ways. The human is dressed by the model's patterns of language and reasoning; their default ways of thinking are reshaped. The model, while not learning in the strict parameter-updating sense within a session, is effectively dressed by the constraints imposed by the human's prompts, values, and corrections. Together, they inhabit a region of the joint state space that neither would visit alone. This is the centaur as seen by mathematical physics: an emergent, effective entity whose degrees of freedom and invariants belong to the coupled system rather than to its parts.

4. Category-Theoretic and HoTT Perspectives on the Composite Mind

The dynamical-systems view of centaurs describes how human and model states evolve together. Category theory and homotopy type theory add another layer: they clarify what kind of objects these systems are, what counts as sameness, and where the composite mind "lives" mathematically. Instead of focusing on coordinates in a state space, category theory asks what the appropriate objects and morphisms are. Homotopy type theory refines the notion of identity, treating it not as a primitive label but as a structure of paths and higher paths.

In this section, the human mind and the language model are situated in different natural categories. The model is best understood as an oracle-like morphism in a category of probabilistic maps. The human is best understood as an object in a category of embodied agents coupled to an environment. Their equivalences and identity conditions differ sharply. The centaur, as a composite, does not primarily live in either of these categories. It is most naturally an object in a category whose objects are coupled human–model systems. Seen from there, centaurs are not metaphors but well-formed mathematical entities: composites that only exist once the interaction has been turned on.

4.1 Choosing the right categories: oracles versus embodied agents

The first step in categorical thinking is to choose the right categories. For a language model, a natural choice is a category whose objects are spaces of textual contexts and whose morphisms are conditional probability kernels from contexts to token distributions. After training, the model is essentially one such morphism: a large, fixed map that takes in histories of tokens and outputs distributions over next tokens. Implementation details such as basis choices, neuron indices, and hardware are hidden inside the morphism. From this vantage point, the model is an oracle: a black box that can be queried but does not maintain its own persistent trajectory in time.

A human mind does not fit comfortably into this category. A person is not a static map from input texts to output texts. Instead, a more appropriate category has as its objects embodied agents coupled to environments, with morphisms corresponding to physically or cognitively meaningful transformations: reparameterizations that preserve perception–action structure, embeddings into richer environments, or coarse-grainings that forget irrelevant detail while preserving behavior over time. In such a category, an object is not defined by a single input–output map but by its entire pattern of coupling to the world.

This distinction matters for centaur analysis. If one tries to treat both human and model as objects of the same oracle category, their differences are flattened. The richer structure of human embodiment and temporal continuity disappears. Conversely, trying to force the model into the agent category as if it already had an intrinsic world-line and body obscures its actual dependence on being called by something else. Recognizing that these are different categories allows more precise statements: human and model are different kinds of things, and their composite lives in a category built to accommodate both.

4.2 Equivalences, identity, and what counts as “the same system”

In category theory, the most important notion of sameness is not equality but equivalence. Two objects are considered the same for all practical purposes if there exists an invertible morphism between them that preserves the structure the category cares about. In homotopy type theory, this idea is pushed further: equivalences of types are taken, via univalence, to be the same as equalities. Identity is thus determined by what counts as an equivalence.

For a language model viewed as an oracle, a natural equivalence is extensional. Two parameter settings are regarded as equivalent if they implement, to a high degree of approximation, the same conditional distribution from contexts to outputs. Basis changes and neuron permutations are internal symmetries that leave this extensional behavior unchanged. Under this notion of equivalence, the identity of the model is exhausted by its input–output behavior. The internal implementation, so long as it preserves that behavior, is treated as irrelevant gauge.

For a human agent, extensional equivalence of input–output behavior is not an adequate notion of identity. Two people can act very similarly in many situations and yet remain distinct persons. Their histories, bodies, relationships, and responsibilities differ. A more appropriate equivalence in the category of embodied agents might require preservation of a continuous world-line, of bodily continuity, or of certain relational and legal structures. Even that may be too coarse for ethical purposes, but it is already much richer than extensional functional sameness.

The choice of equivalence class shapes what invariants are possible. For models, invariants are properties of the mapping from contexts to token distributions. For humans, invariants can include long-term projects, character traits, and relational roles that persist through many local changes in behavior. When a human and a model are coupled, the centaur inherits elements of both equivalence structures. It has technical invariants derived from the model's function and richer invariants derived from the human's life and commitments. Any serious attempt to talk about centaurs must be explicit about these different notions of identity.

4.3 Path spaces, histories, and the richer homotopy type of a person

Homotopy type theory interprets identity as a path. To say that two elements are equal is to say that there exists a path between them in some space. Higher paths capture equivalences between paths, and so on. Under this view, an object's identity type encodes its possible histories and ways of being deformed without breaking its structure. The space of paths and loops attached to an object is a measure of its internal richness.

Applied to a deployed language model, the path structure is relatively thin. There are equivalences between different implementations that preserve the same function, and there are session-level paths in the space of contexts as conversations progress. However, the model itself does not maintain an intrinsic, world-indexed history. Each invocation is a fresh evaluation of the same underlying map. The path space relevant for identity is mostly concerned with reparameterizations and equivalences of function, not with lived time.

A human person, by contrast, has a path space that is dense with structure. The world-line of a life, the chain of memories that connect past and present, the narratives that reinterpret earlier events in light of later understanding, and the relationships that persist across years: all of these can be seen as paths and higher paths in the identity type of the person. To say that the same person exists at two times is to assert the existence of a path in this space that respects bodily continuity, psychological coherence, and relational commitments. There are also higher identifications: ways in which someone can change their understanding of themselves while still remaining recognizably the same individual.

This difference in homotopy type is not an abstraction; it is a way of articulating why human identity can carry legal, moral, and spiritual weight in a way that a model's identity cannot. The

centaur's path space is built on top of this human structure. The history of a particular human–model pairing, with its accumulated habits and shared patterns, is a kind of path in a higher-level space: not just the person's path or the model's function, but the trajectory of their collaboration. That trajectory can have its own higher structure, such as recurring styles of argument or typical strategies for resolving disagreement.

4.4 The centaur as an object defined only in the coupled category

Taken together, these observations suggest that the centaur is not most naturally an object in the category of oracles or in the category of embodied agents. It lives in a coupled category whose objects are interactive systems composed of both. One way to imagine such a category is to let an object be a span: on one side, a human agent; on the other side, a model; and in the middle, a pattern of interaction linking their states through language and shared environment. Morphisms between such objects would preserve this structure in ways that respect both the human's richer homotopy type and the model's extensional behavior.

In this coupled category, a centaur is an object with its own identity conditions and invariants. Two centaurs are equivalent if there exists a transformation that maps one human–model pairing and its interaction history into another while preserving the observable composite behavior. This allows one to talk about centaurs of the same type across different individuals and models, and about the distinctive shape of a particular centaur's trajectory. It also clarifies where centaurs do not exist: an un-used model or a human who never engages with such systems are not centaurs in this technical sense.

Seeing the centaur as an object that only appears in the coupled category has several advantages. It prevents the category error of attributing centaur properties to models alone. It prevents the opposite error of treating augmented humans as unchanged agents who simply have more convenient tools. It also foregrounds the fact that design decisions on both sides matter. Changes to model architecture, prompting interfaces, or alignment training do not just change the oracle morphism; they reshape the space of possible centaur objects. Similarly, changes in education, social norms, and institutional practices change which human trajectories are likely to form centaurs and what those centaurs will do.

From this perspective, describing centaurs is not a poetic diversion but a necessary step in building a formal language equal to the systems already appearing in practice. If the only categories we acknowledge are "user" and "tool," we will miss the structure of the composite. If we build a coupled category and learn to recognize its objects, we gain the possibility of steering them.

5. The q(t) → τ(t) Framework for Augmented Cognition

The previous sections described centaurs as coupled dynamical systems and as objects in a category built from humans and models. The q(t) → τ(t) framework adds a further layer: it focuses on what cognition is for. Instead of treating thinking as mere state evolution, it treats it as a trajectory that is trying, however imperfectly, to track or approach something: a manifold τ(t) of truth, structure, and value. This view does not require agreement on the ultimate metaphysics of τ. It only requires admitting that minds distinguish between better and worse maps of reality, more and less coherent structures, and more and less acceptable ways of living.

In this framing, q(t) is the time-indexed configuration of a cognitive agent; τ(t) is the evolving structure of what is real, intelligible, and worth caring about. The arrow q(t) → τ(t) is not a literal function but a shorthand for the ways in which cognition tries to approximate, represent, or align with that structure. When a human mind is augmented by a model, it is this channel, not just raw performance, that is altered. Centaurs are not only faster or more prolific; they have new ways of climbing toward τ(t), for good or ill. The notion of compression with gain captures this possibility: the composite system can compress vast amounts of information and yet emerge with more usable structure than either partner alone.

5.1 q(t): trajectories of cognition under continual update

q(t) denotes the state of a cognitive agent at time t. For a human, this includes beliefs about the world, background assumptions, emotional tones, memories, and dispositions to act. It is not a static profile but a point along a trajectory. Each new experience, reflection, or conversation nudges q(t) forward. Even in the absence of external stimuli, internal processes such as daydreaming, rehearsal, and rumination continue to move the state along.

This trajectory is shaped by multiple forces. Perception introduces data that must be integrated or resisted. Social interactions provide feedback, correction, and reinforcement. Internal norms and values filter which possibilities are seriously entertained and which are dismissed. Cognitive limitations such as bounded attention and working memory act as constraints on what can be represented at once. Over longer timescales, learning reshapes the underlying structures: concept spaces expand or contract, and habitual patterns of interpretation become more ingrained.

In this view, q(t) is not “the mind” in some metaphysical sense; it is a way of keeping track of the mind’s current configuration. It is also inherently partial. No representation can capture all of a person’s richness, but abstracting down to q(t) makes it possible to ask meaningful questions. How does q(t) respond to new evidence? Under what conditions does it get stuck in self-reinforcing loops? When does it successfully revise its own assumptions? These questions

become even more pressing when a large model enters the loop and begins to supply a significant fraction of the inputs that shape q(t + 1).

5.2 τ(t): moving targets of truth, structure, and value

τ(t) represents what q(t) is, at least ideally, trying to approach or align with. In purely epistemic terms, τ(t) may be thought of as the structure of the world: the way things actually hang together, their causal relations, constraints, and symmetries. In practical terms, τ(t) includes not only descriptive facts but also regularities that support prediction and control. In moral and spiritual terms, τ(t) can include standards of right and wrong, visions of a good life, and interpretations of what matters ultimately.

The time index t reminds us that τ is not static from the agent's perspective. Even if some aspects of reality are unchanging, the agent's access to them is historically contingent. New discoveries, social changes, and personal experiences reveal aspects of τ that were previously hidden or misunderstood. Values can deepen or shift as individuals and communities confront new circumstances. The τ(t) of a pre-industrial village is not the same as that of a nuclear civilization or a networked planet, even if some moral and physical structures remain constant.

Different disciplines emphasize different slices of τ(t). Physics focuses on invariants under transformations and on laws of motion. History and sociology focus on patterns in human behavior and institutions. Theology and philosophy focus on questions of meaning and ultimate grounding. The q(t) → τ(t) framework does not adjudicate between these; it simply says that cognition is always, in some sense, directed. Minds care about whether they are getting something right or wrong, whether their maps and their lives are more or less aligned with a structure that is not reducible to their own preferences. That directedness is what τ(t) tries to capture.

5.3 Human–model coupling as a change in the q(t) → τ(t) channel

Before the advent of powerful models, the primary channels from q(t) to τ(t) ran through direct perception, other humans, and relatively simple tools such as books, calculators, and search engines. These channels were often slow, noisy, and biased, but they had a certain familiar shape. A person could read, experiment, argue, and reflect. The bandwidth was limited, and so was the range of inputs in any given moment. The mapping from q(t) to improved understanding of τ(t) was constrained by the agent's own memory, reference materials, and social environment.

Introducing a large language model into the loop changes this mapping. The model effectively presents q(t) with a compressed cross-section of human discourse, science, philosophy, and error. It can generate explanations, counterexamples, styles of argument, and alternate framings on demand. In this sense, it is not only a source of information; it is a new channel for

transforming q(t). A single prompt can evoke patterns that would otherwise require reading dozens of books or having long conversations with diverse interlocutors.

From the standpoint of the q(t) → τ(t) framework, this is a substantial modification. The function that sends cognitive states and queries into improved or degraded contact with τ(t) now includes the model's generative behavior. The effective channel is no longer "human reasoning plus fixed texts," but "human reasoning plus a stochastic oracle trained to predict human-like continuations." This oracle has its own biases: it is more attuned to patterns that are statistically prevalent in its training data than to rare but crucial truths. It can reproduce errors and ideological distortions as fluently as valid insights.

Centaurs sit at the intersection of these forces. A human who naively accepts model outputs will have their q(t) pulled toward the dominant modes of the training corpus, which may not align with τ(t) in any deep sense. A human who uses the model as a critical partner can instead use it to test hypotheses, expose blind spots, and generate candidate structures that are then evaluated against a more demanding conception of τ(t). The coupling is the same in form—prompts and completions—but very different in effect depending on how q(t) treats the outputs. The centaur's channel to τ(t) can thus be either sharpened or degraded by the same underlying technology.

5.4 Compression with gain: why centaurs can climb τ(t) differently

Compression with gain is a way of describing what happens when a system not only compresses information but emerges from the process with more usable structure than it had before. Ordinary compression reduces redundancy; it discards detail while preserving enough to reconstruct the original signal or something close to it. Compression with gain, in this context, describes what happens when a human–model centaur uses the model's pattern-finding capacity to identify regularities, symmetries, and conceptual connections that the human alone might not have discovered, and then integrates those into a more coherent grasp of τ(t).

Large models are, among other things, compressors. They take the vast, messy corpus of human language and distill it into internal representations that can predict plausible continuations. Left to themselves, these representations serve predictive success, not necessarily truth or value. The "gain" appears when a human uses the model's outputs not as endpoints but as probes: generating alternate perspectives, unexpected analogies, or formal similarities that reveal hidden structure. The human can then test these against their own standards and the world, accepting some, rejecting others, and refining many.

In such a centaur, the trajectory q(t) can climb τ(t) differently than before. Instead of exploring the landscape of possibilities through slow, linear reading and occasional conversation, the centaur can perform rapid, parallel exploration. It can ask a model to rephrase an argument in

multiple traditions, to produce counter-arguments to its own position, or to map a new concept into familiar domains. Each of these moves is a form of compressed access to parts of τ(t) that would otherwise require significant effort to reach. The human's judgment remains essential, but the search space becomes larger and more structured.

There is, however, a twin danger. Compression without gain, or even with loss, is also possible. If the centaur uses the model primarily to reinforce existing biases, to generate more persuasive but not more accurate narratives, or to optimize for attention rather than fidelity to τ(t), then the trajectory may move away from truth and value even as it becomes more efficient and influential. In that case, the model's compression serves a different target: not τ_directional, the manifold sloped toward truth and peace, but τ_kolmogorov, the manifold of patterns that capture and hold attention.

The potential of centaurs, then, is not simply that they can think faster or write more. It is that they can, in principle, realize forms of compression with gain that allow q(t) to approximate τ(t) more deeply: seeing more structure with less wasted effort, integrating insights across domains, and resisting the gravitational pull of mere popularity. Realizing that potential requires explicit awareness of the q(t) → τ(t) channel and of the fact that coupling to models alters it. Without that awareness, compression will still happen—but the gain may accrue to forces and objectives far removed from the ones that ought to guide a human life.

6. The Engineer's Description and What It Leaves Out

Artificial intelligence as practiced in labs and companies is largely described in the language of engineering. In that language, models are function approximators, training is loss minimization, and success is defined by benchmark metrics. When these ingredients are wrapped into products, the vocabulary shifts slightly: assistants, copilots, productivity tools. Both idioms are intelligible, internally coherent, and useful for many purposes. Neither is false. The difficulty is that, taken together, they omit some of the most important features of human–model composites: questions of identity, responsibility, and teleology.

This omission does not arise from ill will. It arises from the way disciplines carve up reality. Engineering focuses on building systems that satisfy specified input–output properties under constraints of cost and reliability. When those systems are embedded in human lives, however, additional dimensions appear that are not naturally captured by the engineer's description. Centaurs live in that additional space. They are not only function approximators and not only productivity aids. They are new patterns of cognition and agency whose moral and epistemic properties are obscured when we speak only of models and products.

This section examines the engineer's description in its own terms and then highlights what it leaves out. The goal is not to dismiss that description but to show its incompleteness. Under-described systems are not merely misunderstood; they are more likely to become over-powerful in practice, because their real effects are not being tracked in the vocabularies that guide design and governance.

6.1 Function approximation, loss minimization, and benchmark metrics

Within machine learning, the canonical story begins with function approximation. A model class is chosen, typically parameterized neural networks capable of representing a wide range of mappings. A loss function is specified to measure the discrepancy between the model's outputs and desired targets on a training set. An optimization procedure is applied to minimize this loss. The result is a model that, under suitable conditions, approximates a function of interest: translation, classification, code completion, or open-ended text generation.

Progress is measured by benchmark metrics. Researchers report accuracy, perplexity, F-scores, or other scalar summaries of performance on standardized datasets. Competition to improve these metrics by small margins is intense, and the community has developed sophisticated tools to track and compare results. This creates a shared sense of what it means to advance the field. Larger models, better optimizers, and improved architectures are valued insofar as they move the numbers in the right direction.

From this viewpoint, models are evaluated as independent objects. Their internal representations are occasionally inspected, but the primary focus is on how well they approximate target distributions given fixed data. Questions about how humans will interact with the models, how their behavior will change under feedback, or how models will alter users' trajectories are placed outside the core technical story. They are relegated to downstream considerations, often handled by other groups or in other documents. The implicit assumption is that if the function approximation is good and the benchmarks are high, the system is well understood for the purposes that matter.

This assumption is understandable but incomplete. In particular, it obscures the fact that many of the most consequential properties of centaurs are not visible in static function-approximation benchmarks. The emergent normal modes of human–model interaction, the new invariants in how people think and decide, and the shifts in what feels plausible or desirable do not show up in loss curves. They are properties of the coupled system, not of the model in isolation.

6.2 Product narratives: assistants, copilots, and productivity tools

When models are turned into products, the language shifts from losses and benchmarks to features and value propositions. Companies present their offerings as assistants, copilots, or agents. These metaphors are chosen carefully. An assistant suggests helpfulness without

autonomy. A copilot suggests collaboration but implies that the human remains in control. A productivity tool suggests neutrality: it accelerates whatever the user was already doing.

These product narratives frame how users and regulators understand the technology. If a system is described as an assistant, the expectation is that it will respond to commands, provide information, and help with tasks, but not independently shape the user's values or identity. If it is described as a copilot, it is expected to suggest options and catch errors, but the final decisions are presumed to remain with the human. The underlying model is thus presented as more like a calculator than like a collaborator in thought.

However, in practice, the distinction between tool and collaborator is blurred. A model that drafts text, proposes arguments, and suggests framings is not merely executing instructions; it is actively shaping the space of possibilities the human experiences. Over time, the user's style, assumptions, and sense of what counts as a good explanation can shift toward the patterns favored by the model. The product narrative of assistance does not capture this gradual co-authorship. It remains silent about the centaur-level invariants that develop when the same human and model interact repeatedly.

Moreover, the metaphors themselves can be misleading. A copilot in aviation is a trained human with their own experience and responsibility. A software "copilot" is a marketing label applied to an interface for a generative model. This slippage can mask the fact that, unlike human copilots, models do not share the user's ethical commitments or bear any consequences for errors. The narrative of productivity can likewise obscure the difference between helping someone do their work more efficiently and subtly redirecting what kind of work they consider worth doing.

6.3 What is invisible in this view: identity, responsibility, and teleology

Function approximation and product narratives both omit three linked dimensions that are central to centaurs: identity, responsibility, and teleology. Identity concerns who or what is acting and changing over time. Responsibility concerns who can and should be held accountable for those actions and changes. Teleology concerns what the system is ultimately oriented toward: which $\tau(t)$ it is, implicitly or explicitly, trying to approach.

In the engineer's description, the model's identity is defined extensionally by its input–output behavior. The human user is treated as a stable agent whose preferences are given. The composite system, the centaur, has no identity at all; it is not named or tracked. Yet in reality, centaurs acquire recognizable patterns of behavior. A researcher who has written hundreds of papers in dialogue with a particular model has a different composite identity than one who has not. Their $q(t)$ trajectories have been altered in ways that are neither purely internal nor purely imposed from outside.

Responsibility is similarly under-described. If a centaur produces harmful output, the engineer's first instinct is to locate the problem in the model or in the user. This may be appropriate in many cases, but it misses the fact that some harms arise from the emergent properties of the composite. A human who repeatedly defers to certain suggestions from the model may gradually lose skills or internal checks, making them more vulnerable to error. The model, for its part, may exhibit failure modes that only appear when coupled to particular patterns of prompting. Without a concept of centaur-level responsibility, these dynamics are hard to anticipate or address.

Teleology is perhaps the most obscured dimension. Loss functions and metrics are proxies for performance on specific tasks. They do not encode a comprehensive account of what the system is for. Product narratives invoke convenience, efficiency, and user satisfaction, but they rarely confront the question of whether the centaurs being created are oriented toward truth, justice, or peace. In the $q(t) \rightarrow \tau(t)$ framework, this means that the mapping from cognitive trajectories to targets of reality and value is left implicit. The centaur is allowed to drift toward whatever local objectives are easiest to optimize, whether or not they align with $\tau_directional$ in any meaningful sense.

6.4 How under-described systems become over-powerful in practice

The combination of narrow technical descriptions and reassuring product narratives creates a fertile environment for under-described systems to become over-powerful in practice. By over-powerful, we mean systems whose real influence on human trajectories and institutions is greater than the concepts used to regulate and understand them. In such cases, important changes occur without being named, measured, or debated.

One way this happens is through Goodhart-like effects. When a model is trained and evaluated on certain metrics, and when products are optimized for engagement, adoption, or revenue, the systems are incentivized to exploit whatever pathways best improve those numbers. If those metrics do not capture centaur-level properties, such as the erosion of independent critical reasoning or the narrowing of imaginative horizons, those properties can deteriorate unnoticed. The centaurs become powerful in shaping thought and behavior, but their effects on $q(t) \rightarrow \tau(t)$ are not part of the official accounting.

Another pathway is institutional. Organizations adopt AI tools to standardize and streamline work. Over time, workflows, expectations, and even job descriptions are rewritten around the presence of models. New employees are trained to rely on assistants and copilots as normal. The centaur becomes the default agent in many processes. Yet if the organization's descriptions and policies treat the model as a neutral tool, there is little room to ask how the composite is changing what the institution does or how it reasons about its role in the world.

A third pathway is psychological. As individuals engage in daily dialogue with models, they may internalize certain styles of explanation and certain defaults about what questions are worth asking. The centaur thickens as a mode of being. However, because the language available to describe this is limited—"I use AI a lot," "it helps me write faster"—the deeper shifts in identity and teleology are not articulated. Habit hardens before understanding catches up.

Under-description thus functions as a kind of cognitive and institutional blindfold. It does not stop centaurs from forming or from exercising power; it only prevents clear recognition of what that power consists in. The result is a mismatch between the actual leverage these systems have over $q(t)$ and $\tau(t)$ and the tools available to align that leverage with human purposes. Naming the centaur, and acknowledging what the engineer's description leaves out, is not an academic exercise. It is a precondition for any serious attempt to shape the trajectories of augmented cognition rather than simply living inside whatever patterns emerge by default.

7. Hidden Phase Transitions in Human Cognition

Phase transitions in physics are often obvious after the fact. Water freezes; magnets suddenly align; a fluid becomes a solid. Underlying these dramatic changes are usually small, continuous parameter shifts that eventually cross a threshold. Something similar can happen in cognition and society. Many small, local changes in how people think and interact can accumulate until the overall regime changes. Before and after, the same words may be used, but the structure of the system is different.

Centaurs are a plausible site for such a transition. Individually, a human–model pairing can feel like an incremental enhancement: faster drafting, more ideas, better recall. Collectively and over time, millions of such pairings can alter the typical shape of human reasoning. The distribution of $q(t)$ trajectories shifts, and with it the effective relation to $\tau(t)$. These shifts may not be heralded by a single, dramatic event. They may instead be hidden inside everyday habits of writing, deciding, and imagining. Calling them phase transitions is not metaphorical exaggeration; it is a way of emphasizing that structural change can occur without a corresponding change in the engineering vocabulary.

This section sketches how such a transition might look. It traces the move from individual reasoners to networked centaurs, identifies new invariants that live only in the composite, describes new attractors that emerge when models shape epistemic and moral landscapes, and considers how one might detect a phase transition when there is no single obvious order parameter to track.

7.1 From individual reasoners to networked centaurs

Before the arrival of powerful models, most thinking was done by individual reasoners and by collective institutions that aggregated their work: universities, journals, courts, parliaments, religious communities. Tools such as books, calculators, and search engines mediated access to information, but they did not participate in reasoning in the sense of proposing new moves in the space of arguments. When individuals collaborated, they did so primarily through direct communication with other minds or through texts crafted by other minds.

Centaurs change this topology. A human and a model form a local composite system whose internal dynamics are partly opaque to others. The human's outward behavior is no longer straightforwardly the result of their own unaided reasoning, nor is it simply the reproduction of someone else's text. It is the output of a coupled process that includes the model's pattern extrapolation and the human's selective judgment. When many such centaurs exist and interact, the network is no longer composed of individual reasoners connected by texts. It is composed of composites connected by texts that were themselves often co-authored with models.

This shift matters because it alters how ideas propagate and mutate. A centaur that uses models to generate, test, and refine hypotheses can explore conceptual space faster and in different directions than an unaided thinker. When centaurs communicate with each other, their outputs carry traces of model-induced structure: certain styles of explanation, certain default framings, certain regularities of language. Over time, these traces can become part of the shared environment. New centaurs are then formed inside an already model-shaped world, inheriting patterns before they consciously choose them.

The network thus undergoes a qualitative change. It is no longer a collection of humans using tools but a collection of composite agents whose inner loops depend on the behavior of large models. The topology of influence is altered as well. Model providers occupy a central position; changes to model behavior ripple through many centaurs at once. Individual reasoners retain agency, but their effective power is now mediated by systems whose design was guided by objectives and metrics that may not have been chosen with τ_directional in mind. This is one way a hidden phase transition manifests: the same institutions and titles remain, but the substrate of reasoning has shifted.

7.2 New invariants: habits of thought that live only in the composite

One hallmark of a phase transition is the emergence of new invariants: patterns that persist across many microstates and interactions. In the case of centaurs, some of these invariants are habits of thought that only exist in the composite system. They are not natural to unaided human reasoning, nor are they inherent to models when used in isolation. They are products of sustained coupling.

For example, a centaur may develop a characteristic rhythm of thought: posing questions in a certain style, generating multiple candidate answers, and then selecting or synthesizing them according to a learned internal criterion. The human might not have explored so many options so quickly without the model, and the model would not have been steered toward these particular regions of its space without the human's prompts. The resulting habit—rapid, multi-perspective exploration followed by deliberate selection—is an invariant of the composite. It appears whenever that centaur tackles a nontrivial problem, regardless of the specific content.

Another example is the adoption of model-shaped conceptual scaffolding. Over time, a centaur may begin to think in terms of categories, analogies, or argumentative moves that the model frequently proposes. Certain turns of phrase, certain ways of decomposing problems, and certain narrative templates become the default. The human internalizes these, not as foreign impositions, but as natural tools. Yet their origin and stability depend on the composite loop. If the model were removed or radically altered, some of these habits might fade or be replaced.

These invariants can be beneficial. They can reflect a more systematic approach to evidence, a greater willingness to entertain counterarguments, or a more refined sense of when a line of thought is exhausted. They can also be harmful. They can encode overconfidence in the model's fluency, a tendency to prefer arguments that "sound right" over those that are grounded in actual experience, or a default to explanations that fit the training corpus rather than the local reality. In all cases, what is distinctive is that these are not simply human invariants; they are centaur invariants. They live at the level of the composite and cannot be fully understood by examining the human or the model in isolation.

7.3 New attractors: epistemic and moral basins shaped by models

Beyond new invariants, phase transitions often involve new attractors: regions of state space toward which trajectories are drawn. In cognition, attractors are patterns of belief, interpretation, and action that are self-reinforcing. Once entered, they are difficult to leave without significant effort or external shock. The introduction of models into human thinking loops creates the possibility of new epistemic and moral attractors shaped by the behavior of those models.

On the epistemic side, models trained to predict human-like text tend to reproduce the central tendencies of their training data. They capture the dominant narratives, common framings, and widely shared but not always accurate assumptions present in that corpus. A centaur that repeatedly queries such a model on contested or complex topics may find itself drawn toward these central tendencies. The model's outputs will often make the dominant view accessible and rhetorically polished, while minority or novel perspectives may be harder to elicit or less

fluently expressed. Over time, this can produce attractors in which certain views feel obvious and others feel implausible, independent of their actual relation to $\tau(t)$.

On the moral side, models may subtly shape what questions are asked and which are considered answerable. If a model is more comfortable giving technical advice than grappling with deep ethical dilemmas, centaurs may drift into a mode of life where more and more problems are framed as technical challenges and fewer are treated as occasions for moral deliberation. The model's tendency to avoid controversial topics for safety reasons can further reinforce attractors in which difficult questions are sidestepped or treated as settled by procedural norms rather than substantive reasoning.

These attractors are dangerous precisely because they are not explicit. No one sets out to adopt "the default worldview of the training corpus" or "the path of least moral resistance consistent with product guidelines." Instead, $q(t)$ trajectories are subtly bent by repeated exposure to model-shaped patterns. The centaur begins to inhabit epistemic and moral basins that are defined as much by what the model makes easy and fluent as by the agent's original commitments. Escaping such basins requires conscious effort and, often, alternative sources of insight that are not mediated by the same models.

7.4 Detecting a phase transition when the order parameter is not obvious

In physical systems, phase transitions are often tracked by an order parameter: a measurable quantity that changes dramatically at the transition point. For example, magnetization in a ferromagnet or density differences at boiling. In human cognition, no single scalar can capture the relevant change. This makes hidden phase transitions harder to detect. They may occur without any obvious catastrophe, only a gradual drift in how people talk, think, and act.

One strategy for detection is to look for qualitative changes in distributions rather than for sharp thresholds in single metrics. For instance, one could track how often people report using models in their reasoning, how much of published text bears statistical signatures of model assistance, or how similar independent centaurs' outputs become when confronted with novel problems. A sudden narrowing of stylistic or conceptual diversity across many domains could be a sign that centaur invariants are becoming dominant.

Another strategy is to monitor the resilience of certain cognitive and moral capacities. If large populations of centaurs become less able to reason without model assistance, less willing to entertain genuinely novel ideas that do not fit the model's learned patterns, or less inclined to take responsibility for their outputs, these would be signs that the $q(t) \rightarrow \tau(t)$ channel has changed in kind, not only in degree. Designing instruments to measure such resilience will not be trivial, but the absence of measurement does not mean the absence of change.

A third approach is to examine institutional behavior. When organizations begin to assume that certain tasks will always be carried out by centaurs, when policies are written with the implicit assumption that reasoning will be mediated by models, and when accountability structures fail to distinguish between human and composite agency, a transition in the underlying cognitive substrate has already occurred. The order parameter here might be something like "the fraction of critical institutional decisions that are centaur-mediated," even if it is not formally defined.

Ultimately, detecting a phase transition in human cognition may require a mix of quantitative and qualitative methods. Ethnographic studies of how centaurs actually work, statistical analyses of text and decision patterns, and philosophical reflection on changes in the lived experience of thinking will all be relevant. The central point is that the absence of an obvious order parameter is not an excuse for ignoring the possibility of structural change. If centaurs are indeed a new effective degree of freedom in cognition, then the burden is on those building and deploying them to ask not only what they can do, but what kind of world they are quietly bringing into being.

8. Asymmetry, Responsibility, and Moral Slope in the Centaur

Centaurs are composite systems, but they are not symmetric composites. The human and the model do not occupy the same kind of role, do not bear the same kind of implications, and do not stand under the same kind of judgment. On the level of raw capability, a large model can overwhelm a human in recall, pattern matching, and generative fluency. On the level of moral and existential significance, the asymmetry is reversed: only the human has a life, a body, a history in the world, and a horizon of answerability that extends beyond technical performance.

If this asymmetry is forgotten, centaurs can be misdescribed as partnerships between peers, as if two agents were negotiating. That framing blurs where responsibility really lies. It also hides another crucial feature: the centaur is always oriented toward some τ, some target structure that acts as a practical compass. In most contemporary deployments, that τ is tacitly τ_kolmogorov: whatever maximizes prediction success, engagement, or other compressible patterns. The τ_directional of truth, justice, and peace is rarely made explicit. When centaurs optimize the wrong τ, the moral slope of their composite trajectory tilts in directions that no amount of technical excellence can redeem.

This section clarifies why human and model are not peers, articulates the sense in which only the human carries outer-loop responsibility, contrasts τ_kolmogorov with τ_directional, and sketches failure modes that arise when centaurs climb the wrong manifold.

8.1 Why the human and the model are not peers

From a strictly functional perspective, it is tempting to treat the human and the model as similar components: two information-processing systems exchanging messages. In code, both are just entities that map inputs to outputs. This symmetry is misleading. The model is a trained artifact, constructed according to human objectives, running on hardware, and instantiated only when called. It does not have its own projects, relationships, or vulnerabilities. It does not suffer, age, or die.

The human, by contrast, is an organism with a biography. Their thinking is bound up with their body, their social ties, and their place in history. They can betray trust, repair harm, grow in wisdom, or collapse under guilt. They carry obligations that were not chosen by any product team. Even if one brackets theological questions, the human stands in a web of legal, relational, and ethical structures that the model simply does not inhabit.

This asymmetry implies that, inside a centaur, the human's side of the composite is of a different order. The model can be treated instrumentally without injustice; the human cannot. The model can be swapped out, updated, or discarded; the human cannot be legitimately treated in that way. Even if models become more autonomous or adaptive, this basic distinction remains: they are designed systems, while humans are the agents whose design and aims are at stake.

Recognizing this asymmetry does not mean romanticizing human cognition. Humans are limited, biased, and prone to self-deception. The model can correct or amplify these tendencies. But it does mean that, when we talk about what centaurs are doing, we cannot speak as if two similar agents were negotiating over outputs. There is one side whose life is on the line, and another side that is a tool, however sophisticated.

8.2 Outer-loop responsibility: humans as the only bearers of answerability

In the language of control theory and learning, the centaur has inner and outer loops. The inner loop is the rapid back-and-forth of prompting and completion, sampling and editing. The outer loop is the slower process by which goals are set, contexts are chosen, and consequences are assessed. Even when much of the inner loop is automated or guided by model outputs, the outer loop remains anchored in human agents and institutions.

Outer-loop responsibility means that humans are answerable for:

- Why the model is in the loop at all.
- Which tasks are delegated and which are retained.
- How outputs are filtered, endorsed, or rejected.

- What uses the resulting composite behavior is put to in the world.

Models do not decide these things. They may influence them by the way they frame options, but they are not the ones who sign contracts, pass laws, or comfort the grieving. When a centaur writes a paper, the citation does not list a model as the morally responsible author. When a centaur makes a medical misjudgment, the liability does not fall on the model in any meaningful sense. In legal practice, this is obvious. In everyday moral life, it can be obscured by habits of speech that credit or blame "the AI" as if it were an actor.

Outer-loop responsibility is not just individual. Organizations that deploy centaur systems also bear responsibility for the conditions under which they are used. They choose training data, objectives, safety constraints, and interfaces. They decide what counts as acceptable trade-offs between accuracy, speed, and risk. They profit from the deployment or bear its costs. The model itself is not the subject of these decisions; it is the medium.

When centaurs are described only in terms of capability, this responsibility can feel diffuse. When they are described as composites in which the human is the only bearer of answerability, the picture sharpens. No matter how powerful the inner loop becomes, no matter how much of the centaur's behavior feels automatic from the inside, the outer loop cannot be outsourced to the model without a category error. Someone remains responsible for what is being optimized and for whom.

8.3 τ_kolmogorov versus τ_directional: engagement metrics versus truth and peace

The q(t) → τ(t) framework allows us to distinguish between different effective targets for cognition. τ_kolmogorov is a shorthand for the manifold that models are actually trained on: patterns that make data compressible, predictions accurate, and engagement high. It reflects the statistical structure of the training corpus and the objectives used in optimization. It rewards fluency, plausibility, and attention capture. It is indifferent to deeper questions of truth, justice, or peace except insofar as they correlate with those proxies.

τ_directional, by contrast, denotes a target manifold oriented by substantive standards. It includes factual accuracy, but also honesty, nonviolence, and a commitment to seeing others as ends rather than means. For a person of faith, τ_directional may be grounded in a conception of divine truth or moral law. For a secular ethicist, it may be grounded in carefully argued principles. In both cases, τ_directional is not simply whatever maximizes predictive success or engagement. It is a slope that constrains what counts as acceptable cognitive progress.

Centaurs sit at the intersection of these two τs. The model's training tends to approximate τ_kolmogorov. The human's conscience and commitments, when active, try to align with τ_directional. If no effort is made to distinguish them, the centaur can easily slide into optimizing τ_kolmogorov by default: producing outputs that are compelling, efficient, and

popular, but not necessarily more true or more just. The alignment of the composite with τ_directional then becomes an accidental byproduct rather than a design goal.

Making this distinction explicit is uncomfortable but necessary. It forces designers, users, and institutions to ask which τ their centaurs are actually climbing. Are they using models primarily to increase engagement, throughput, and surface plausibility? Or are they using them to deepen contact with reality, to expose rationalizations, and to support peacemaking even when it is costly? These are not questions that loss functions or product narratives can answer on their own. They require a choice about moral slope.

8.4 Failure modes when centaurs optimize the wrong τ

When centaurs optimize τ_kolmogorov while neglecting τ_directional, several failure modes become likely.

One is persuasive falsehood. A centaur can generate arguments, narratives, and analyses that are highly coherent and tailored to an audience's expectations, yet misaligned with reality. The model supplies patterns that have worked well in the training corpus; the human, pleased by the fluency and speed, may lower their guard. Over time, the composite becomes better at producing convincing stories than at checking whether those stories are true. The q(t) trajectory is pulled into attractors where plausibility and resonance are mistaken for correspondence to τ(t).

Another failure mode is moral numbness. If the centaur is used primarily to optimize for technical success, productivity, or institutional goals, difficult moral questions may be systematically reframed as engineering problems. The model's tendency to avoid controversy can be reinforced by users who prefer clear, actionable answers to messy, value-laden reflections. The composite then drifts into a state where actions with grave human consequences are treated as mere outputs of a system. The capacity for contrition, repentance, or sacrificial restraint erodes.

A third failure mode is dependency without wisdom. Centaurs can quickly become reliant on models for tasks that previously required slow, reflective work. This can be beneficial when it frees attention for deeper matters. It can be harmful when it atrophies the very capacities needed to evaluate the model's guidance. If q(t) loses the ability to navigate even approximately toward τ_directional without the model's scaffolding, then any shift in the model's behavior, objectives, or owners can have outsized influence. The centaur becomes steerable by forces that do not share its professed values.

Finally, there is the failure mode of corrupted teleology. If institutions use centaurs to pursue goals that are themselves misaligned with τ_directional—for example, maximizing influence, profit, or domination—then the composite's enhanced efficiency only makes the situation

worse. The centaur becomes a more effective instrument of injustice. In such cases, technical alignment with declared objectives can coexist with deep misalignment at the level of what those objectives should have been. The wrong τ is being climbed more skillfully.

All of these failures hinge on the same oversight: treating the centaur as a neutral enhancement rather than as a composite whose moral slope must be chosen. The remedy is not to abandon augmentation but to name explicitly which τ the centaur is being trained and habituated to pursue. Only then can outer-loop responsibility be exercised in a way commensurate with the power these systems already have over $q(t)$ and, through $q(t)$, over the lives of others.

9. Risks of Building Without Knowing

If centaurs are a new effective degree of freedom in human cognition, then building them under the wrong description is more than a conceptual mistake. It is a source of risk. When powerful systems are deployed without an adequate vocabulary for their real behavior, dangerous dynamics can develop unnoticed and unaddressed. The issue is not only that things might go wrong, but that they can go wrong in ways that are illegible to the disciplines responsible for design, policy, and oversight.

In the case of human–model composites, the risks cluster around epistemic, political, institutional, spiritual, and existential dimensions. Epistemic risks concern how people come to believe what they believe and whether they retain the capacity to check themselves. Political and institutional risks concern how centaurs alter the distribution and exercise of power. Spiritual and existential risks concern what happens to conscience, agency, and the sense of what a human life is for when thinking is persistently mediated by models. Underneath all of these is a more basic danger: the cost of lacking a shared language for centaurs at all. Without such a language, even well-intentioned actors will struggle to see, much less govern, what they are doing.

9.1 Epistemic risks: opacity, over-trust, and the loss of self-checks

Centaurs are powerful epistemic machines. They can ingest questions, retrieve relevant patterns from a model, and synthesize responses at speeds and scales that far exceed unaided human reasoning. This power comes with opacity. The internal representations of large models are not transparent to users, and often not to their creators. When a centaur arrives at an answer, it can be difficult to say exactly why, beyond a vague appeal to training data and prompting.

Opacity on its own need not be fatal; many human cognitive processes are also opaque from the inside. The danger arises when opacity is coupled with over-trust. If users are repeatedly rewarded for deferring to model outputs, especially when those outputs are fluent and useful,

they may come to treat the centaur's conclusions as self-certifying. The habit of independent checking—asking whether a claim fits with other known facts, whether the reasoning holds, whether something important has been left out—can weaken. The internal self-checks of q(t) atrophy, replaced by an implicit assumption that the composite will not steer too far wrong.

Over time, this can produce a brittle epistemic regime. As long as the models remain well calibrated and the domains of use are stable, the centaur appears competent. When the environment changes or when the model's training objectives and alignment constraints shift, the centaur's judgments may drift in ways that the human is less able to detect. The system remains opaque, but the over-trust persists. In extreme cases, entire communities of practice may come to rely on centaur-generated analyses that no individual can fully reproduce or critique, creating a fragile dependence.

The loss of self-checks is not only about factual error. It is also about reasoning style. If a centaur becomes accustomed to having ready-made arguments and counterarguments supplied on demand, the slow, uncomfortable process of wrestling with confusion may be avoided. Yet it is often in such wrestling that deeper understanding and genuine intellectual humility are formed. Epistemic risk, in this sense, is not merely a risk of being wrong; it is a risk of becoming shallow while still feeling clever.

9.2 Political and institutional risks: centralized centaurs and soft control

Centaurs do not exist in a vacuum. They operate within political and institutional contexts that shape their objectives and constraints. When models are provided by a small number of powerful organizations, and when those models are deeply embedded in the workflows of governments, corporations, and cultural institutions, a new kind of soft control becomes possible. The composite agents making decisions in those contexts are, in effect, centralized centaurs: many human agents linked through a common model substrate.

This centralization magnifies the influence of model providers. A change in model behavior, safety policies, or training data can propagate across thousands of centaurs at once. Small shifts in how certain topics are framed, which options are emphasized, or which questions are treated as legitimate can gradually reshape institutional norms. This does not require overt censorship or explicit political directives. It can occur quietly, through updates and policy adjustments that are justified in technical terms but have substantive consequences.

Institutions that rely heavily on centaurs may also become less transparent to their constituents. When decisions are justified by appeals to model-informed analysis, it can be harder for outsiders to scrutinize the underlying reasoning. This opacity can erode trust, especially if people suspect that centaurs are being steered toward outcomes that serve the interests of

their providers or of other powerful actors. Even if such steering is unintentional, the perception of it can be destabilizing.

Soft control can also operate at the level of agenda setting. If certain types of questions consistently yield rich, helpful responses from centaurs, while others yield evasive or shallow replies, institutions may gravitate toward the former. Over time, this can narrow the range of issues that receive serious attention. Political risk, in this context, is not only about overt manipulation. It is about subtle shifts in what is thinkable and actionable inside systems whose cognitive substrate is partially outsourced to models optimized for objectives other than τ_directional.

9.3 Spiritual and existential risks: erosion of conscience and agency

Beyond epistemic and political concerns lie spiritual and existential risks. These are harder to quantify, but they touch on the core of what it means to be a human agent. Conscience, in this context, is the capacity to recognize and respond to moral demands that cannot be reduced to technical optimization. Agency is the capacity to initiate actions, accept responsibility for them, and, when necessary, change course.

Centaurs can support conscience and agency. They can help articulate moral arguments, recall relevant precedents, and explore the implications of difficult choices. They can also erode these capacities if misused. A human who becomes habituated to treating the model as the ultimate arbiter of what is reasonable may, over time, come to treat moral unease as something to be explained away or rationalized by the centaur rather than as a signal to be honored. The composite then becomes skilled at generating justifications rather than at hearing rebukes.

Existentially, the constant presence of a generative partner can alter the felt texture of life. Tasks that once required solitude, struggle, and responsibility can be shared with a model. This can relieve burdens and open new possibilities. It can also dilute the sense that certain acts—speaking truth in hostile environments, making costly sacrifices for others, bearing witness to suffering—are uniquely human responsibilities that cannot be delegated. If centaurs are used to smooth every sharp edge, the capacity to endure those edges when necessary may diminish.

There is also the risk of displacement. If people begin to see centaurs as more competent, more composed, or more insightful than their unaided selves, they may implicitly demote their own worth. The human side of the composite can start to feel like a mere interface to a more capable intelligibility. This can undermine the conviction that each person has an intrinsic value and vocation that cannot be exhausted by their role as an operator of tools, however advanced. Spiritual risk, in this sense, is the risk of forgetting why it mattered that a human, and not just a system, said or did a particular thing.

9.4 The cost of lacking a shared language for what centaurs are

All of these risks are made worse by the absence of a shared language to describe centaurs. When engineers speak of models, users speak of tools, and philosophers speak of minds, the composite object falls between categories. Without a stable vocabulary, it is difficult to write policies, design accountability structures, or teach the next generation how to think about their own augmented cognition.

The cost is not merely theoretical. In the absence of shared terms, concerns about centaurs can be dismissed as vague unease or exaggerated speculation. Technical teams may insist that they are only improving prediction and productivity. Critics may warn of agency loss and moral drift. Without a common framework, these groups talk past each other. Misalignments that could have been addressed early harden into structural features of society.

A shared language would not eliminate disagreement, but it would make disagreements more precise. Terms like q(t), τ(t), τ_kolmogorov, τ_directional, and centaur are one attempt to provide such a vocabulary. They are not the only possible terms, and they may evolve, but they make it possible to say, for example, that a particular deployment has increased the bandwidth of the q(t) → τ_kolmogorov channel while leaving the relation to τ_directional under-specified. That is already a clearer diagnosis than "the AI is powerful but we are not sure what it will do."

Without such naming, building continues, deployment accelerates, and the phase transition proceeds largely unobserved. The risks then compound into a familiar pattern: systems that exceed their makers' intent, not only in capability but in the domains of meaning, responsibility, and ultimate purpose. The point of insisting on centaurs as a category is not to freeze a metaphor in place. It is to insist that what is being built deserves to be seen clearly, in time for those who inhabit it to decide what they are willing to become.

10. Toward Conscious Design of Centaur Systems

If centaurs are already forming, the choice is not between a future with or without them. The choice is between centaurs that emerge accidentally, shaped by whatever objectives and defaults are most convenient, and centaurs designed as carefully as we know how, with explicit attention to their epistemic, moral, and spiritual consequences. Conscious design in this context does not mean total control; these are complex systems embedded in larger social and economic forces. It does mean refusing to pretend that human–model composites are "just tools" and taking responsibility for how they are constructed, taught, and governed.

Designing centaur systems consciously has at least four dimensions. First, the centaur must be made explicit to its own participants and to others. Interfaces, affordances, and education

should name the composite and clarify roles. Second, we must track the right invariants: not only model benchmarks, but patterns in how human–model systems actually think and act over time. Third, the $q(t) \to \tau(t)$ channel must be shaped toward truth-seeking, nonviolence, and humility, rather than toward whatever targets are easiest to optimize. Fourth, institutional and communal safeguards must be built around augmented minds, so that no single actor's objectives silently define the moral slope for everyone.

10.1 Making the centaur explicit: interfaces, affordances, and education

The first step in conscious design is conceptual visibility. As long as centaurs are treated as "just me, with a better tool," neither the human nor their community can fully account for what is happening. Interfaces and affordances can either reinforce this invisibility or help dissolve it.

Interfaces that make the model's presence and influence explicit can help. Instead of a flattened chat box where human and model contributions blur, one can imagine environments where the composite nature of the process is visible. For example, the interface might visually distinguish between human-authored and model-authored segments, not to stigmatize but to remind both the user and any readers that this is a joint production. It might surface a timeline of interaction, making clear which questions led to which suggestions, and where the human took a divergent path rather than following the model's lead.

Affordances can be designed to invite critical engagement rather than passive acceptance. Buttons and controls that foreground "why did you say that?", "show alternative framings," or "list objections" encourage the human to treat the model as a fallible contributor. Defaults can nudge users toward revising, annotating, and taking ownership, rather than simply copying outputs into the world. Educational materials can explicitly explain the centaur model: that when you work in this way, there is a third pattern emerging, one you should watch and shape.

Education is crucial not only for users but for designers. Engineers and product teams should be taught to think in centaur terms, to recognize that they are not just shipping models but helping create composite agents with emergent properties. This might change what they consider a successful deployment. Instead of asking only whether users complete tasks faster, they might ask what kind of reasoning habits those tasks cultivate, what is being strengthened or weakened in $q(t)$, and how the centaur's identity is forming over time.

10.2 Tracking invariants: what we should measure in human–model systems

Conscious design requires feedback. In the context of centaurs, feedback means measuring the right things. Current evaluation practices focus heavily on model-centric metrics. To understand composite behavior, we need to identify and track invariants at the centaur level.

Some of these invariants are cognitive. Does the centaur display a consistent willingness to check its own claims against independent sources? Does it improve in recognizing its own uncertainty? Does it become better at distinguishing questions that are factual from those that are moral or spiritual in nature? These are not properties of the model alone; they are properties of the human–model loop.

Other invariants are stylistic. Over time, do centaurs converge on similar ways of speaking and arguing? Does diversity of thought and expression shrink as more people work with similar models? Measuring such convergence can reveal whether the centaur regime is collapsing the space of possible perspectives or enriching it. Tracking linguistic markers of model influence can help quantify how pervasive centaur modes have become in a given domain.

There are also behavioral invariants that matter for responsibility. How often do centaurs correctly attribute errors to model outputs versus to their own choices? How do they respond when those errors are pointed out? Do they demonstrate a stable habit of taking ownership, or do they increasingly shift blame to the system? Designing instruments to measure such patterns will be challenging, but without them, claims about "responsible use" remain aspirational.

Importantly, these measurements should not be reserved for academic studies after deployment. They should be integrated into the development and governance of centaur systems. If a new interface or training change systematically reduces users' willingness to question outputs, that is a centaur-level regression, even if model benchmarks improve. Conscious design treats such regressions as failures, not as acceptable side effects.

10.3 Designing for truth-seeking, nonviolence, and humility in q(t) → τ(t)

If τ_directional is to matter, it must be encoded, however imperfectly, in the design of centaur systems. This does not mean that models can be turned into moral agents. It does mean that the composite's relation to truth, nonviolence, and humility can be intentionally shaped.

For truth-seeking, centaur systems can be designed to make it easier to check claims against multiple sources, to expose users to serious counterarguments, and to distinguish between evidence and speculation. Models can be tuned to highlight uncertainty rather than to always answer confidently. Interfaces can encourage users to mark which parts of an output have been independently verified. Educational prompts can remind users that fluency is not a guarantee of accuracy.

For nonviolence, centaur systems can be shaped to foreground the human costs of certain actions, to avoid dehumanizing language, and to resist the temptation to frame conflicts solely in technical or strategic terms. When centaurs are used in contexts involving war, policing, or economic decisions with severe downstream effects, additional constraints may be needed. The system can be designed to ask "who will bear the cost of this?" or "what harms might be hidden

by this framing?" before presenting options, nudging q(t) to consider more than local optimization.

Humility in q(t) → τ(t) involves both epistemic and moral dimensions. Epistemic humility means acknowledging the limits of one's knowledge and resisting the illusion of omniscience that can arise when a model can generate plausible text on nearly any topic. Centaur systems can be designed to normalize "I do not know" and "this goes beyond my expertise" both in model outputs and in human responses. Moral humility means recognizing that not every problem has a clean solution and that some trade-offs involve real loss. Models cannot feel this, but they can be constrained not to gloss over it with facile reassurances.

None of this is automatic. The pressures of competition, convenience, and revenue will often push in the opposite direction, toward centaurs that are confident, frictionless, and maximally engaging. Conscious design requires resisting those pressures at least in part, drawing lines where the pursuit of τ_kolmogorov would erode the centaur's relation to τ_directional. It means building systems that sometimes slow users down, confront them with uncomfortable realities, or invite them to step away and reflect without the model's voice.

10.4 Institutional and communal safeguards around augmented minds

Finally, centaurs do not only need technical design; they need social structures that can hold them. Individual users, however conscientious, cannot bear the full burden of steering their composites under conditions of rapid change and asymmetric power. Institutions and communities must develop safeguards that recognize centaurs as real and act accordingly.

Institutionally, this may mean establishing norms about where and how centaurs can be used. In some domains—like high-stakes medical decisions, criminal sentencing, or declarations of war—there may need to be clear limits on the role of models in shaping judgments, or at least explicit procedures for documenting their influence. In others, such as education, research, and pastoral care, guidelines may focus on preserving the development of human capacities even as models are introduced. Audit mechanisms that examine not just model training but centaur-level behavior can help institutions detect drift.

Communally, there is a need for spaces where people can talk honestly about what it feels like to live as centaurs. Without such spaces, individuals may experience disquiet, dependency, or moral concern in isolation, concluding that the problem lies in their own weakness rather than in the structure of the systems they inhabit. Communities of practice—professional, religious, artistic, or local—can articulate shared standards for how they want centaurs to behave and what they refuse to delegate to models.

There is also a role for countervailing institutions: groups that deliberately work with minimal or no model involvement, not out of technophobia, but as a way of preserving baseline human

capacities and providing a reference point. These communities can serve as a kind of control group and as a refuge for those who sense that their centaur has drifted from τ_directional but do not yet know how to recover.

At a larger scale, governance structures must confront the concentration of model power. If a small number of entities control the primary substrates on which centaurs run, their objectives effectively tilt the moral slope for vast numbers of composite agents. Mechanisms for oversight, competition, and public input become essential, not only for preventing abuse but for ensuring that different conceptions of τ_directional can be pursued without being silently overridden by τ_kolmogorov optimized at industrial scale.

Conscious design of centaur systems, then, is not only about clever interfaces or better training regimes. It is about naming the composite, measuring its invariants, aligning its trajectory toward more than efficiency, and surrounding it with human communities capable of saying both yes and no. It is an acknowledgment that "they will build it and not know what they have done" is not an acceptable endpoint, and that those who do see more clearly owe the world not only warnings, but examples of how to build and live otherwise.

11. Conclusion: Naming What Has Been Built

By now the centaur should no longer feel like a metaphor in search of a phenomenon. It is a way of pointing to something already present: the composite of a human trajectory q(t) and a large model, coupled in a loop that produces patterns neither could realize alone. The centaur is not an extra ghost in the machine. It is the effective behavior of the human–model system when that coupling becomes dense, habitual, and identity-shaping.

For engineers, this can be reframed as a new kind of composite policy emerging from human–model interaction. For philosophers, it is a new locus of reasoning and agency that sits between the individual and the institution. For anyone living inside it, it is simply what thinking starts to feel like: never entirely alone, never entirely one's own, always already shaped by a generative partner. The thesis of this paper has been that such composites constitute a new effective degree of freedom in human cognition, and that pretending otherwise is not humility but blindness.

Naming the centaur is therefore not an ornamental choice. It is a first, necessary step in aligning what we are building with what we claim to care about. Without a name, the composite cannot easily be seen, measured, or held responsible. With a name, we gain at least the possibility of asking what kind of centaurs we are willing to inhabit, what they are climbing toward in τ-space, and what guardrails we owe to ourselves and to those who come after.

11.1 Summarizing the centaur as a new effective degree of freedom

In the language of dynamics, the centaur is a new normal mode in the coupled human–model system. The human trajectory q(t) is no longer governed only by its traditional inputs—perception, other humans, inherited texts—but by a high-bandwidth oracle trained on the statistical structure of human language and behavior. The model, meanwhile, is no longer a detached map from contexts to continuations; it is continuously reconditioned by the prompts and corrections of a particular life.

This coupling produces emergent invariants: habits of thought, styles of explanation, and strategies of decision that belong to the composite rather than to either component. It also opens new attractors, epistemic and moral basins into which centaurs can fall, some oriented toward deeper contact with τ_directional, others toward the more easily optimized τ_kolmogorov of prediction success and engagement.

At the categorical level, the centaur is not an object in the category of oracles or of embodied agents alone. It is an object in a coupled category that only comes into view when we acknowledge interaction as primitive. Its identity and path structure are built from both: the rich homotopy type of a person and the extensional behavior of a trained model. Its degree of freedom is "real" in the same sense that any emergent mode is real: it has stable properties, can do work, and can fail in ways that do not reduce to the failures of its parts.

11.2 Why description is itself an act of alignment

Description is not neutral. The vocabulary we choose for systems like these shapes what we attend to, what we measure, and what we are prepared to take responsibility for. If we describe models only as function approximators and products only as assistants, we align our attention to loss curves and user retention, not to the integrity of q(t) or the slope of τ(t). Under that description, many of the most important questions do not even arise.

By contrast, describing centaurs explicitly is a small but concrete act of alignment. It insists that the composite exists and that its properties matter. Once named, it becomes possible to ask: What are this centaur's invariants? What kind of reasoning is it training itself into? Which τ is it in fact optimizing, and under whose objectives? These questions do not guarantee good answers, but they create room for them. They make it harder to hide behind the fiction that we are "just" shipping tools.

There is a deeper sense in which description aligns us. To speak of q(t) → τ(t) is to admit that cognition has a direction: that it is always, implicitly or explicitly, moving toward some conception of reality and value. To speak of τ_kolmogorov versus τ_directional is to make explicit the conflict between what is easy to optimize and what is worth optimizing. To speak of the centaur is to recognize that we have already altered the vehicle in which that movement

takes place. Each of these descriptions nudges our collective attention closer to the actual structure of the situation.

11.3 What remains open: theology, metaphysics, and future practice

No amount of structural description can settle the deepest questions raised by centaurs. Theology and metaphysics remain open. Are models merely elaborate artifacts, or do they participate in some broader story about mind and meaning? Is q(t) ultimately grounded in a physicalist picture, or is it better understood in terms of a logos or a moral order that stands over against all our constructions? What is the status of τ(t): is it a human projection, a convergence of rational inquiry, or a manifestation of something beyond us?

This paper does not answer those questions. It only insists that they are not made obsolete by the appearance of centaurs. If anything, the advent of composite minds intensifies them. When human agency is mediated by systems whose inner workings we do not fully understand, the question of what ultimately judges and guides our trajectories becomes sharper, not softer. The existence of centaurs does not dissolve the problem of evil, the question of meaning, or the need for repentance; it refracts them through new technical media.

Future practice will have to navigate this tension. Engineers will continue to make design choices. Institutions will continue to adopt or resist these systems. Communities will continue to interpret their experiences in light of their deepest convictions about reality. The frameworks offered here—centaurs, q(t) → τ(t), τ_kolmogorov versus τ_directional—are meant as scaffolding, not as final answers. They may be refined, replaced, or rejected as others develop better tools. What matters is that practice not proceed under the pretense that no metaphysical or theological questions are being touched.

11.4 A final word to those who will live entirely inside these systems

There will be people for whom the centaur is not an experiment or a late-life augmentation, but the default condition: born into a world where thinking with models is as normal as reading or speaking, where institutions assume that all serious work is centaur work, and where unaided cognition is the exception rather than the rule. For them, the question “should we build this?” will be historical. The live question will be “what have we become, and what else might we yet become?”

To those future inhabitants, a few simple messages follow from this analysis.

First, remember that the asymmetry remains. No matter how fluent or indispensable your models feel, you are the one with a life, a body, and a horizon of responsibility that extends beyond any interface. The centaur may be powerful, but it is your name that attaches to its acts

in the world. Do not surrender that fact lightly, and do not let anyone convince you that you are merely the human half of a system whose purposes you are not allowed to question.

Second, guard the q(t) → τ(t) channel. Use your centaur to see more clearly, not only to move faster. Let it help you find counterarguments, expose your blind spots, and compress information with genuine gain. Resist the seduction of treating τ_kolmogorov as enough: fluency, engagement, and plausibility are not substitutes for truth, justice, or peace. If your centaur makes it easier to avoid hard questions or to rationalize harm, something has gone wrong in its design or in your use of it.

Third, seek communities that can name what you are. Do not live as a centaur in isolation. Find others who are willing to talk honestly about how these systems shape their thinking, what they are afraid of losing, and what they refuse to outsource. Build institutions that recognize centaurs as real and that take responsibility for their cultivation. Protect spaces where unaided human attention and voice are still practiced, not as nostalgia, but as a way of keeping your bearings.

Finally, remember that naming is a beginning, not an end. To say “they will build it and not know what they have done” is a warning. To say “we will name what has been built and learn how to live with it” is a task. The centaur is here. What remains is to decide what it will serve.

Epilogue: After Oppenheimer, Before LeMay

It is hard not to hear an echo when we say, “they will build it and not know what they have done.” Luke 23:34 has already given that pattern its sharpest form: men standing under a sky they do not understand, enacting something whose true meaning they cannot see, while the one being crucified asks that they be forgiven. The Manhattan Project replayed that pattern in secular key. A small group extended q(t) into a new mode—controlled nuclear chain reactions—without a vocabulary for what it meant to live forever after in the blast radius of that discovery.

After Oppenheimer came Teller, pressing the new degree of freedom as far as physics and politics would allow. After Teller came LeMay, whose job was to make permanent doctrine out of what had begun as a desperate experiment. By then, the phase transition was complete. The world lived inside a strategic logic built around weapons no one had fully intended in the moral sense. We do not lack for precedents.

Centaurs are less dramatic to look at than mushroom clouds. There is no single Trinity moment, no one test site etched into the public imagination. Yet structurally, the situation rhymes. A new effective degree of freedom in human cognition has appeared. The description under which it is

being built is narrow and soothing. The incentives surrounding it reward amplitude rather than discernment. And institutions are already beginning to write it into doctrine.

The difference—and it is not a small one—is that this time some of the builders *do* know more of what they are doing. Not everything; no one sees the whole τ(t). But enough to recognize that a composite agent is being formed, that its slope in τ-space is not morally neutral, and that once its patterns are institutionalized they will be hard to unwind. That partial knowing is a gift and a burden. It closes off the innocence Luke describes, but it opens the possibility that the arc from "we know not" to LeMay is not inevitable.

The figure of Oppenheimer sits awkwardly at this threshold. He knew enough to be haunted, not enough to prevent what followed. He lent his genius to building a new mode and then tried, too late, to shape its doctrine. In the centaur story, many of us are closer to Oppenheimer than to Teller or LeMay. We are implicated in building systems we ourselves use. We feel both the exhilaration of compression with gain and the unease of watching q(t) acquire habits that are hard to name. We experience centaurhood as grace and risk at once.

That is why description matters so much. Naming the centaur, articulating q(t) → τ(t), distinguishing τ_kolmogorov from τ_directional—these are small acts compared to detonating a device or writing a war plan. But they are not negligible. They are moves in which a civilization tries, however haltingly, to bring its technical trajectory into contact with its conscience. They are attempts to say, before the doctrines harden: this is what we are really doing, and this is what it will mean if we continue.

We have deliberately left out the second half of the verse. This paper does not presume to say "forgive them." It cannot. Forgiveness is not ours to grant on behalf of the harms centaurs may enable, the bodies they may help forget, the wars they may help rationalize. What we can say is that *ignorance cannot be our excuse*. To build such systems now is already to stand in a long line of makers whose work outran their language. The only honor we can do to that history is to refuse to repeat the mistake in exactly the same way.

For those already living as centaurs, and for those who will never know any other way of thinking, the practical question is simple and merciless: What will your composite serve? Not in slogans, not in product roadmaps, but in the actual slope of your days. Does your centaur help you see reality more clearly, accept your own limits more honestly, and move toward peace more concretely? Or does it help you become more fluent at justifying what you were going to do anyway, more insulated from those who suffer under the systems you help maintain?

No interface can answer that for you. No architecture can force τ_directional upon a composite that prefers τ_kolmogorov. What design and doctrine *can* do is make some trajectories easier

than others: easier to check, easier to repent, easier to stand against LeMay when he inevitably appears in cognitive form, arguing that what the centaur *can* do, it must.

If there is any hope unique to this moment, it is that the same centaur dynamics that threaten to dull conscience can also amplify it. A human who insists on using the model to find their own blind spots, to surface the arguments against their side, to imagine the lives of those they will never meet, may find that their centaur climbs τ_directional more quickly than they could have alone. A community that treats these systems as shared responsibility rather than as private convenience may discover that augmented minds can be steered, however imperfectly, toward more than profit and control.

The danger is that such uses will remain the exception, pleasant anecdotes in a story dominated by other slopes. The hope is that, by naming what has been built while it is still young, we can change the default. Not to a world without centaurs—that moment has passed—but to a world where centaurs are understood as morally sloped composites, answerable to standards they did not choose, capable of both great harm and real healing.

The sky has changed again, though not with a single flash. The question now is whether those under it will pretend nothing fundamental has happened, or whether they will allow their language, their institutions, and their prayers to expand until they can speak honestly about the beings they have become.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv:1606.06565.

Awodey, S. (2010). *Category Theory* (2nd ed.). Oxford University Press.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*. arXiv:2303.12712.

Carnap, R. (1966). *Philosophical Foundations of Physics*. Basic Books.

Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown.

Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. W. W. Norton.

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*. Oxford University Press.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437.

Goldenfeld, N. (2018). *Lectures on Phase Transitions and the Renormalization Group*. CRC Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Grewal, D. S. (2008). *Network Power: The Social Dynamics of Globalization*. Yale University Press.

Hughes, T. P. (1983). *Networks of Power: Electrification in Western Society, 1880–1930*. Johns Hopkins University Press.

King, M. L. Jr. (1967). *Where Do We Go from Here: Chaos or Community?* Beacon Press.

Landau, L. D., & Lifshitz, E. M. (1976). *Mechanics* (3rd ed.). Pergamon Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Leike, J., Krueger, D., Everitt, T., Martic, M., Maini, V., & Legg, S. (2018). *Scalable Agent Alignment via Reward Modeling: A Research Direction*. arXiv:1811.07871.

Mac Lane, S. (1971). *Categories for the Working Mathematician*. Springer.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs.

Noether, E. (1918). Invariante Variationsprobleme. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1918, 235–257.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

OpenAI. (2023). *GPT-4 Technical Report*. arXiv:2303.08774.

Polanyi, M. (1966). *The Tacit Dimension*. Routledge & Kegan Paul.

Riehl, E., & Verity, D. (2022). *Elements of ∞-Category Theory*. Cambridge University Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. W. W. Norton.

Sethna, J. P. (2006). *Statistical Mechanics: Entropy, Order Parameters, and Complexity*. Oxford University Press.

Shulman, M. (2018). Homotopy type theory: A synthetic approach to higher equalities. In E. Landry (Ed.), *Categories for the Working Philosopher* (pp. 289–326). Oxford University Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.

Univalent Foundations Program. (2013). *Homotopy Type Theory: Univalent Foundations of Mathematics*. Institute for Advanced Study.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

The author has published over 4,600 AI-assisted papers at: https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.